On the Complexities of Federating Research Data Infrastructures

Atif Latif, Fidan Limani & Klaus Tochtermann[†]

ZBW – Leibniz Information Centre for Economics, 24105 Kiel/Neuer Jungfernstieg 21, 20354 Hamburg, Germany

Keywords: Metadata; Harvesting; Repository federation; Research data infrastructures

Citation: Latif, A., Limani, F., Tochtermann, K.: On the complexities of federating research data infrastructures. Data Intelligence 3(1), 79-87 (2021). doi: 10.1162/dint_a_00080

ABSTRACT

Federated Research Data Infrastructures aim to provide seamless access to research data along with services to facilitate the researchers in performing their data management tasks. During our research on Open Science (OS), we have built cross-disciplinary federated infrastructures for different types of (open) digital resources: Open Data (OD), Open Educational Resources (OER), and open access documents. In each case, our approach targeted only the resource "metadata". Based on this experience, we identified some challenges that we had to overcome again and again: lack of (i) harvesters, (ii) common metadata models and (iii) metadata mapping tools. In this paper, we report on the challenges we faced in the federated infrastructure projects we were involved with. We structure the report based on the three challenges listed above.

1. INTRODUCTION

The advancement in data intensive science has powered a generation of various digital scientific artefacts like research data, source code, scripts, workflows and algorithms. The recent 2018 European Commission report on *Turning FAIR into Reality* [1] indicated the importance of these artefacts and stressed their development in compliance to a FAIR Digital Objects ecosystem[①]. One of the majorly discussed artefacts for fostering the Open Science (OS) convergence is research data and their management. At a policy level, research data have become a focal point in the European Union OS policy processes, which states the

[†] Corresponding author: Klaus Tochtermann (Email: k.tochtermann@zbw.eu; ORCID: 0000-0003-2471-2697).

[①] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects exchange of research data within scientific disciplines to create added value for the progress of science, innovation, transparency and reproducibility, and finally quality of scientific results [2]. However, the main challenges for science policy and infrastructure projects are to first educate the scientific communities about data openness and further develop research practices and prerequisites that data publishers need to adhere to. Moreover, it is also very important to comply with good scientific standards such as FAIR principles [3] to make the research data discoverable, accessible, citable and interoperable for society potential reuse. In order to reap the benefits of research data, funding agencies and many (trans)national initiatives such as the European Open Science Cloud (EOSC)[②] and the GO FAIR Initiative[③] are already pushing for a set of criteria that research data need to abide by.

To achieve the practical realization of these envisioned criteria, developing services for research data management is important to ease researchers in their data-related activities. On the other hand, to facilitate heterogeneous communities of research disciplines, a seamless integration and availability of these services to relevant infrastructure is also getting pertinent and critical.

One of the widely accepted solutions is to develop Federated Research Data Infrastructures (FRDI) [4] for data related service federation. Generally, such an infrastructure is one where a range of distributed services—focused on the actual research requirements/needs—are coordinated comprehensively, with the aim to provide potentially seamless access to research data and services.

In the midst of the federation of RDI initiatives, we see that communities usually lack established research management practices, including adopted metadata standards and services tailored to the specific research lifecycle. During our research on OS, we have built cross-disciplinary federated infrastructures for different types of open digital resources: Open Data (OD), Open Educational Resources (OER), and open access documents. The three shortcomings we had to overcome repeatedly was the lack of (i) harvesters, (ii) common metadata models, and (iii) metadata mapping tools. In this paper, we highlight these limitations of federation by sharing the experience from these three initiatives.

2. CROSS DISCIPLINARY INFRASTRUCTURES

In this section, we give a brief overview of three RDI projects that we worked on, which provide a cross-disciplinary federator service for different types of open digital resources.

1) Generic Research Data Infrastructure (GeRDI)[④] was funded by the German Research Foundation and it carried on from November 1, 2016 to October 31, 2019. GeRDI provided a generic, sustainable and open software connecting long-tailed heterogeneous research data repositories to enable

[②] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[③] https://www.go-fair.org/go-fair-initiative/

[④] https://www.gerdi-project.eu/

multidisciplinary and FAIR research data management [5]. The software is based on common standards and was developed in close collaboration with various research communities to ensure a best match to the requirements of different disciplines. All project results, in particular software, a central search index, a microservice architecture and other services, along with training support and business model, are public and can be reused as a contribution to federated research data projects like the EOSC.

2) EduArc[5] funded by the German Federal Ministry of Education and Research (BMBF) is scheduled for 3 years (October 1, 2018 – March 31, 2022). It is an RDI for cross-university reuse of digital learning materials (OER). The project develops a tried and tested design concept for distributed learning infrastructures with which digital educational resources and other study-relevant information are federated. It investigates the technical, didactic, and organizational conditions for the success of an educational architecture that arises from networking the digital infrastructure of universities and the interaction of state, public, and private actors. Moreover, it brings together decentralized systems via open standards and interfaces and is open to integrate future content providers and users.

3) MOVING[6] targeted mining and provision of open digital resources (multimedia video lectures and open scholarly documents) for the relevant communities. This project was funded by the European Union's Horizon 2020 research and innovation program from April 1, 2016 to March 31, 2019. MOVING is an innovative training platform that enables people from all sectors to improve their information literacy by training them on how to use, choose, reflect, and evaluate data mining methods in their research activities.

3. EXPERIENCES AND CHALLENGES

Often for each repository that is targeted to be connected to a federated infrastructure such as EOSC or the National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur, NFDI), a tailored harvester that meets the specifics of the access interface or the metadata schema is needed. Once the metadata has been harvested, it has to be mapped on a common metadata model of the federated infrastructure. Even though metadata standards such as DataCite[7] do exist, it is always a "negotiation process" to define such a metadata model as it cannot cover all disciplinary components. Finally, a mapping technology solution to map the repository metadata to the common metadata model is needed. A key challenge here is a mapping as loss-free as possible, which is almost never achievable if metadata from different disciplines is to be mapped to the metadata model. This situation leads to significant implementation efforts and for a number of repositories to be connected to a federated infrastructure, we would need generic and tailored harvesters and mapping tools.

[5] https://learninglab.uni-due.de/forschung/projekte/eduarc-digitale-bildungsarchitekturen

[6] http://moving-project.eu/description/

[7] https://datacite.org/

To put these observations into perspective, we next discuss the common challenges we faced—citing concrete examples and facts—in the projects mentioned above.

3.1 Metadata Harvesting

The first step in populating the RDI typically starts with the harvesting of resources, during which resource metadata is collected. Different resource collections adopt different technical solutions for the publication process. For an RDI, this typically implies implementing different means of harvesting for the targeted resources. In the case of GeRDI, we conducted metadata harvesting via different interfaces. These interfaces included the following cases: a) standard but very rare occurrence of dedicated interface built upon Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) for harvesting; b) some with more generic interfaces built as Application Programming Interface (APIs with different data structures); c) interfaces based on resource repository solutions, such as KIT Data Manager and DSpace; d) Git-related interfaces used by communities as repositories storage solutions; and e) one of the least favorable harvesting options is the conventional website interfaces that are only to be harvested by screen scraping.

Similarly for the EduArc project, the main challenge was that all data sources had different Web-portal structures and not all of the data sources had standard APIs that would allow harvesting of the metadata. Therefore, building a focused harvester for each repository that did not have an API was the only way forward. Similarly, to the GeRDI project, due to different and, at times without a given structure for data sources, we had to hard code the harvesters to extract the metadata. In the case of the project MOVING, we used a focused Web domain crawler to harvest specific Web domains. Profoundly, a search engine-based Web crawler was used to collect topic-relevant Web pages for the platform. In the project, we used Google custom search API[®], which supports searches in the whole Web and specified subsets thereof (i.e., within specific domains). The main challenge in using it was that each API call returns 10 Web page links. As a result, there is a maximum limit of 1,000 links per day due to the rate limit.

In summary, we claim that a standardized harvesting interface would have been ideal, but the user requirements also dictated the resource collections to harvest, and there were no options to dictate or change the harvesting interface for the repositories hosting those resources. However, planning and provision of standardized harvesting interface (OAI-PMH, for example) in upcoming research data infrastructure projects will be one of the go-to-solutions. ResourceSync [6], a follow-up to OAI-PMH is another emerging initiative that not only targets the resource metadata for synchronization but also the resources themselves. While aware of it, we did not encounter it in our projects because there was no repository that we harvested had this specification implemented.

3.2 Common Metadata Model

The next challenge encountered in RDI projects is that of a metadata model to represent all harvested metadata. Research communities that generate research artifacts adopt different research practices, and, as

[®] https://developers.google.com/custom-search/v1/overview a result, various metadata models to represent these artefacts. Thus, the diversity faced is high and includes a broad range of cases, from communities that have few or no metadata descriptions of their resources, to those that have an abundance of (often disciplinary) such descriptions. We have experienced this challenge in all of the three projects that we report on next.

In the case of the GeRDI project, we dealt with resources from nine different communities that ranged from humanities and social sciences, life sciences to natural sciences, with specific research disciplines such as alpine environment, microscopy and bioinformatics, digital humanities, and hydrology. As expected, the metadata model was used across these areas stretched from established generic (DublinCore, DataCite, DCAT, etc.) or disciplinary metadata standards (DDI, SDMX, Genome Metadata, DIF—Directory Interchange Format, etc.), to cases where such standards were not adopted at all. Keeping this in context, we approached the metadata schema design in an incremental way by starting with the generic metadata elements to represent as many of the communities as possible. In doing so, we supported the core services of the RDI, which the users would be able to test and use earlier in the project. Afterwards, based on a community-led prioritization process, we started including metadata elements from the individual communities as part of the disciplinary part of the schema. All the while, we were trying to identify potential metadata reuse, especially for the case where different communities use a conceptually similar/same metadata, but have (slightly) different terms for it.

Similarly, in EduArc, different German OER repositories use different metadata models. For example, the Learning Object Metadata (LOM) [7] standard and Learning Resource Metadata Initiative (LRMI) [8] are used in some cases, while many other OER repositories do not follow any metadata standards at all and rely on in-house metadata representations. For EduArc, we chose the LOM standard to design the common data model. As the MOVING platform provides access to a large variety of documents coming from different data sources, the main challenge was in designing a model for the heterogeneous harvested metadata to include information in the index as much as possible. Therefore, as with GeRDI, we designed our model and tried to include most of the information. However, it was evident that in EduArc some information could be missed since it is hard to include all the fields of the repositories in our common data model.

As seen from these three cases, when it comes to a common metadata model in an RDI, one critical question and challenge we face is: *"how to specify a balanced metadata model broad enough to represent different disciplines, yet expressive (or deep) enough to represent all the requirements of the individual communities"*. A broad model like DataCite, which translates to a narrower or minimal set of metadata, generalizes well at the research data infrastructure level. On the other hand, supporting the majority (if not all) of its services, an expressive or disciplinary model like DDI better addresses the specific requirements of individual communities. Another approach we explored in this context is that of the component-based metadata schemas [9]. Such an approach requires the definition of independent (metadata) components that describe certain aspects and subsequently become part of a "catalog". In an RDI project, one should select only the components that address the metadata requirements of the project by creating a metadata profile.

3.3 Metadata Mapping

Once we decide on a metadata model for the research data infrastructure resources, we need to map the harvested metadata from the original sources to the already chosen RDI model (*cf*. Section 3.2). While the mapping is straightforward for the more common metadata elements, (there is never a debate on how to map a *title* element, for example), it is often challenging to find the best metadata element for the more disciplinary elements in the source data. There is no universally agreed understanding on the semantics of metadata elements, and thus research communities often debate (extensively) what an element in their data means and how it should be represented in the metadata model of the RDI. A simple example is (miss) using metadata fields to provide as many descriptions as possible. Take for example, the *description* of a data set: Who is to say the amount of information one should provide in it? How do we deal with established metadata practices of providing detailed, disciplinary metadata through such general elements in a schema? This situation often arises when communities need certain metadata elements for their research practice, but the adopted model does not support them.

In the case of EduArc, each data source has a different Web-portal structure, and hence the harvested metadata are different for each data source. Because we used the LOM standard for designing the common data model, we needed to design a dedicated mapper for each data source's metadata. Furthermore, not all harvested data are mapped into the common data model because these unmapped fields are not included in LOM. Thus, we lost some harvested information due to the unavailability of discipline specific metadata in the model. Similarly, in MOVING, based on the challenge of designing the common data model, some information was not included in the resulting mapped records because missed information did not have a field in the adopted model.

In general, metadata normalization is a necessary activity to try and narrow the domain of values for the different metadata elements in order to provide better services to RDI users. However, due to lack of standard (metadata) mapping options, usually it concedes to the situation where a handful of data information is subjected to loss because the adopted model does not support the certain metadata elements and has to be dealt with categorically.

4. FUTURE RESEARCH AND CONCLUSION

Based on the challenges faced in our RDI projects, we see few paths for future research that could contribute to making OS a reality. To create transparency in this sense, along with existing standard tools and initiatives, registries offer a promising approach. This has been already manifested in other areas, such as the registry for research data repositories like re3data[⑨]. Similarly, registries for harvesters, metadata models [10] and metadata mapping tools (disciplinary or generic) can be reused according to the OS principles. This practice would not only help to create the necessary overview, but also further establish a sharing culture.

[⑨] https://www.re3data.org/

However, it is not enough just to register existing tools and common metadata models. Rather, information about the interrelationships between the tools is also required at a higher level of understanding. Concretely, it should be ensured that the information about which harvester could map which metadata to which (common) metadata model is also preserved.

Based on this, future research should not only focus on the identification and abstraction of common features of different harvesting, mapping tools and common data models, but also on the relationships or dependencies between these three entities. The "knowledge base" can serve as a basis for formal specifications that describe *what* the tools do, but not *how* they do it. Since relationships between entities are to be modelled, semantic technologies, such as Resource Description Framework (RDF), are suitable for a formal description. For example, simple RDF triples might express that harvested metadata (subject) is compliant with a (predicate) specific mapping tool (object) and that a mapping tool (subject) maps onto (predicate) common metadata model (object). Such an approach would allow these triples to be linked to the FAIR Digital Object (FDO) specification; in particular it would be possible to express that metadata [of the FDO specification] is compliant with a specific mapping tool. If we pursue this idea further, we could add an "infrastructure perspective" to the current FDO specification, which models rather than the "data perspective".

Additionally, such semantic representations can be used for the requirements analysis of a generic software framework. It becomes more complex when we think of a step that transforms a specification into a design, which in turn is transformed into an implementation. If it were possible to make this development chain as error-free as possible, harvesters and mapping tools could be developed in the future in a largely, if not completely in the automated manner.

In summary, federated RDIs provide seamless integration of and access to research data management services to support researchers in data intensive tasks. However, the varying community research practices and the heterogeneity present in the area of research data management do make the provision of optimal services and data federation a challenging task. To avoid duplication of work in the future and to facilitate the development of such infrastructures, new design principles for harvesting and mapping tools but also common metadata models are required. We see great potential here for challenging research in the future.

AUTHOR CONTRIBUTION

A. Latif (a.latif@zbw.eu) contributed in conceptualization of paper idea and writing of Sections 1, 2 and 3 with contribution to Section 4; F. Limani (f.limani@zbw.eu) contributed in writing Sections 1, 2 and 3 with contribution to Section 4; K. Tochtermann (k.tochtermann@zbw.eu) conceived the paper idea and contributed in Sections 2 and 3 along with writing of Section 4.

REFERENCES

[1] Collins, S., et al.: Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. Available at: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf. Accessed 30 October 2020

[2] Open Science. European Commission research and innovation. Available at: https://ec.europa.eu/info/sites/info/files/research_and_innovation/knowledge_publications_tools_and_data/documents/ec_rtd_factsheet-open-science_2019.pdf. Accessed 30 October 2020

[3] Wilkinson, M., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3, Article No. 160018 (2016)

[4] Goldstein, S.: The evolving landscape of Federated Research Data Infrastructures. Available at: http://doi.org/10.5281/zenodo.1064730. Accessed 30 October 2020

[5] Latif, A., Limani, F., Tochtermann, K.: A generic research data infrastructure for long tail research data management. Data Science Journal 18(1), 17 (2019)

[6] van de Sompel, H. Overview of ResourceSync. Available at: https://www.niso.org/standards-committees/resourcesync. Accessed 17 December 2020

[7] RISK, U.: Draft standard for learning object metadata. IEEE Report number: 1484.12.4 (2002)

[8] Waters, J.K.: Sifting the data: The learning resource metadata initiative has a complicated name but a simple purpose: To make Web searches more useful for students and teachers. Technological Horizons in Education 40(1), 15–18 (2013)

[9] Broeder, D., et al.: Standardizing a component metadata infrastructure. In: LREC 2012: The 8th International Conference on Language Resources and Evaluation, pp. 1387–1390 (2012)

[10] Metadata Standards Directory WG. Available at: https://www.rd-alliance.org/groups/metadata-standards-directory-working-group.html. Accessed 30 October 2020

AUTHOR BIOGRAPHY

Atif Latif received his PhD degree in Computer Science with a focus on linked Open Data (OD) research from Graz University of Technology, Austria in 2011. Dr. Latif was affiliated with the Institute of Knowledge Management and Know-Center, Austria's COMET Competence Center for Knowledge Management. His main research areas are linked OD, OS and digital libraries. Since 2012, he has been associated with the Leibniz Information Centre for Economics (ZBW) where he investigated on solutions to apply semantic and linked data technologies in digital library settings. Currently, he is researching on applications of FAIR principle, metadata standards and data management practices in the domain of digital data infrastructures and OS.
ORCID: 0000-0003-3085-3031

Fidan Limani, with a background in computer science and information systems, has been engaged with (national) research data infrastructure projects at Leibniz Information Centre for Economics since 2017. This includes research data management aspects and service implementation for different research communities, with foci on analysis and implementation of metadata standards, conception and implementation of automatic metadata generation, automatic metadata linking to standard bibliographic data, and so on. Another part of his research includes the application of semantic Web/linked data technologies as integration means for different scholarly research deliverables into (digital) library environments. He previously worked as a research and teaching assistant at the computer science department of the South East European University in Macedonia for 10 years.
ORCID: 0000-0002-5835-2784

Klaus Tochterman has been director of the ZBW - Leibniz Information Centre for Economics in Kiel and Hamburg (Germany) since 2010. He also holds a full professor position for Digital Information Infrastructures in the Computer Science Department at Christian-Albrechts-University Kiel (Germany). His current research focus is on research data infrastructures and OS. Klaus Tochtermann has repeatedly held guest professorships abroad, such as at the University of St. Gallen (Switzerland) or the Universiti Teknologi MARA (UiTM) in Kuala Lumpur (Malaysia). In December 2020, he was elected as member of the Board of Directors of the EOSC Association.
ORCID: 0000-0003-2471-2697